基于伪三维卷积神经网络的手势姿态估计 *

张宏源 [1]，袁家政 [2†]，刘宏哲 [1]，原春锋 [3]，王雪峤 [1]，邓智方 [1]

(1. 北京联合大学 北京市信息服务工程重点实验室，北京 100101; 2. 北京开放大学，北京 100081; 3. 中国科学院自动化研究所 模式识别国家重点实验室，北京 100190)

摘　要：大多数现有的基于深度学习的手势姿态估计方法都使用标准三维卷积神经网络提取三维特征，估计手部关节坐标。这种方法提取的特征缺乏手部的多尺度信息，限制了手势姿态估计的精度;另外，由于三维卷积神经网络巨大的计算成本和内存需求，这些方法常难以满足实时性要求。为了克服这些缺点，提出以空间滤波器和深度滤波器级联的方式模拟三维卷积，减少网络参数量。同时，在各个尺度上提取手势姿态特征并加以整合，充分利用手势的三维信息。实验表明，该方法能有效提高手势姿态估计精度，减小模型尺寸，且在具有单块 GPU 的计算机上能以超过 119 fps 的速度运行。

关键词：手势姿态估计；伪三维卷积神经网络；三维特征；深度图像；深度学习

中图分类号：TP391.41　　doi: 10.19734/j.issn.1001-3695.2018.09.0772

Hand pose estimation using pseudo-3D convolutional neural network

Zhang Hongyuan[1], Yuan Jiazheng[2†], Liu Hongzhe[1], Yuan Chunfeng[3], Wang Xueqiao[1], Deng Zhifang[1]

(1. *Beijing Key Laboratory of Information Service Engineering*, *Beijing Union University*, *Beijing* 100101, *China*; 2. *Beijing Open University*, *Beijing* 100081, *China*; 3. *National Laboratory of Pattern Recognition*, *Institute of Automation*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: Most of the existing deep learning-based methods for hand pose estimation use a standard three-dimension convolutional neural network (3D-CNN) to extract 3D features and estimate the 3D coordinates of hand joints. The features extracted by these methods lack the multi-scale information of the hand, which limits the accuracy of hand pose estimation. In addition, due to the huge computational cost and memory requirements of the 3D CNN, these methods are often difficult to meet the real-time requirement. To overcome these weaknesses, the proposed method uses a spatial filter and a depth filter to simulate 3D convolutions, which reduces the amount of parameters. The proposed method extracts and integrates features at various scales, making full use of the 3D information of hand pose. Experiments show that the proposed method can improve estimation accuracy, reduce model size, and run at over 119fps on a standard computer with a single GPU.

Key words: hand pose estimation; pseudo-3d convolutional neural network; 3d features; depth image; deep learning

0　引言

基于视觉的手势姿态估计研究近年来取得了显著地进展，作为人机交互的核心技术之一，该技术为用户提供了一种自然地交互方式。由于深度图像可以有效解决单目 RGB 输入中存在的复杂背景干扰等问题，手势姿态估计任务几乎完全转为仅使用深度数据作为输入[1~6]。其次，深度学习改变了视觉问题的解决方式，深度神经网络的使用已经成为手势姿态估计方法中的常态[7~9]。

在众多基于深度神经网络的姿态估计的方法中，深度图常被视为二维图像，输入二维卷积神经网络（convolutional neural network，CNN）中，输出三维关节位置[9,10]、手部模型参数[11]或热图[12]。直观上来说，由于缺乏 3D 空间信息，2D CNN 提取的基于图像的特征并不适用于 3D 手势姿态估计。鉴于此，最近有几种基于 3D CNN 的方法被陆续提出[1,13,14]，然而这些方法只是简单的应用 3D CNN 提取特征，并未充分利用三维信息，同时三维网络的训练需要巨大的计算成本，相比于 2D CNN，模型大小也几乎增加了一倍，为了达到实时性的要求，只能使用较浅的网络结构，这使姿态估计的效果大打折扣。

最近，针对 3D CNN 巨大的计算成本和内存需求问题，文献[15]提出了一种新的网络结构，称为伪三维残差网络（pseudo-3D residual networks，P3D ResNet），这种创新的模块设计在保证准确率的前提下，大幅减小了模型尺寸。文献[16]提出了一种新型的“堆栈式沙漏”网络用于人体姿态估计任务，该设计提取和合并不同尺度下的人体姿态特征，从而显著提升了姿态估计的精度。本文的工作受此启发，提出了一种基于伪三维卷积神经网络的手势姿态估计方法，整体网络结构如图 1 所示。首先将手势姿态的深度图编码为 3D 体积表示并将手部区域从体积表示中分割出来，将其馈送到由基础伪三维残差模块组成的完整网络中，最终输出手部关节的空间坐标。本文方法的优点可以概括如下：

收稿日期：2018-09-02；**修回日期：**2018-11-06　**基金项目：**国家自然科学基金资助项目（61571045）；北京成像技术高精尖创新中心项目（BAICIT-2016002）；北京市教委科技计划一般项目（KM201811417002）；北京联合大学研究生资助项目，

作者简介：张宏源（1993-），男，河南郑州人，硕士研究生，主要研究方向为深度学习、图像处理；袁家政（1971-），男（通信作者），湖南湘潭人，教授，博导，博士，主要研究方向为视觉计算(jiazheng@buu.edu.cn)；刘宏哲（1971-），女，河北保定人，教授，硕导，博士，主要研究方向为数字图像处理；原春锋（1981-），女，山东烟台人，副研究员，博士，主要研究方向为模式识别；王雪峤（1986-），女，讲师，博士，主要研究方向为模式识别；邓智方（1992-），男，河南安阳人，硕士研究生，主要研究方向为深度学习、图像处理．

a）使用改进的手势姿态体积表示方法，训练简单的 CNN 获得更准确的手部区域，去除无效区域的影响；

b）使用伪三维卷积替代标准三维卷积，大幅减小模型尺寸，加快手势姿态估计速度；

c）使用三维“沙漏”结构网络，提取并融合手势姿态多尺度特征，充分利用三维信息，提高手势姿态估计精度。

1 相关研究

1.1 基于深度图像的手势姿态估计

从深度图像中进行手势姿态估计的方法可分为模型生成方法、数据驱动方法和混合法。模型生成方法通常预定义一个手部模型，通过最小化损失函数使手部模型与输入的深度图像相匹配。常见的优化方法是迭代最近点（iterative closest point，ICP）[17]、粒子群优化（particle swam optimization，PSO）[18]或者两者的组合方法[4]。由于这些方法通常需要使用时态信息，因此更依赖于手部模型的初始化，在进行姿态估计时误差也更容易累积。

数据驱动方法直接从输入的深度图中定位手部关节点。受人体姿态估计领域内方法[19]的启发，文献[20,21]使用基于随机森林的方法及其改进方法作为判别模型，获得了准确而快速的性能。然而，受手工设计特征的限制，基于随机森林的方法目前难以超越基于卷积神经网络的姿态估计方法。本文的工作与基于 CNN 的数据驱动方法有关。文献[7]首先提出通过 CNN 估计每个手部关节的 2D 热图，从而定位手部关节点。文献[22,23]提出了一种 2D 区域集合网络（region ensemble network，REN）用来精确估计关节点的三维坐标。文献[16]提出了一种新型的“堆栈式沙漏”网络（Stacked Hourglass Networks，SHN），通过提取并整合各个尺度上图像特征，精确估计二维关节点坐标，并在人体姿态估计领域内取得了较大的成功。文献[1]创新性地将 3D CNN 引入到手势姿态估计任务中来，利用深度图中的三维信息直接估计手部关节点的三维坐标。文献[13]采用多任务级联网络将 2D 关节点检测与 3D 姿态估计两阶段相结合，同时利用深度图的二维和三维信息，实现手势姿态估计。这些方法均使用简单的 3D CNN 来提取手势姿态特征，但并未充分利用深度图中各个尺度的信息。本文的方法利用 SHN 与 3D CNN 两种方法的优势，从多个尺度上提取并整合 3D 特征进行手势姿态估计。

混合法由模型生成和数据驱动两阶段方法结合而来。文献[24]训练了一个由多级网络构成的反馈回路，其中包含进行初始姿态估计的辨别网络，进行姿态合成的生成网络和通过多次迭代改善姿态估计的姿态更新网络。文献[25]使用了两个具有共享潜在空间的深度生成模型，并通过训练鉴别器来估计被遮挡的部分手势姿态。本文的工作专注于进行单阶段的一次性完整手势姿态估计，从而更有效地利用手部关节点之间的潜在相关性。

1.2 伪三维卷积神经网络

3D CNN 已被成功地应用于从深度图和 CAD 模型等数据中提取 3D 特征，进行三维场景重建[26]、三维物体检测[27]、物体识别[28]等任务。然而较浅的 3D CNN 难以获得有效的特征，训练深层 3D CNN 则需要高昂的计算成本和内存需求。针对这些问题，文献[15]中提出的 P3D ResNet 使用空间和时间卷积滤波器组合模拟时空卷积滤波器，这种组合可以被看做为伪三维 CNN，在提升视频分析效果的同时，大幅减少模型参数，缩短模型训练时间。这项工作成功应用于视频相关的多种任务中，但是他们并没有关注 3D 手势姿态估计。本文工作使用空间和深度卷积滤波器模拟三维卷积，提取手势姿态 3D 特征的同时，减少神经网络模型参数，使之满足实际应用场景中的实时性要求，实施细节将在 3.2 小节进行介绍。

2 方法概述

为了充分利用手势姿态深度图中各个尺度的三维信息，加快手势姿态估计速度，本文提出一种新的伪三维“堆栈式沙漏”网络，沙漏网络有助于提取多尺度特征，伪三维结构设计则能有效降低网络训练所需的计算成本和空间需求。首先，含有手势姿态的单张深度图经过体积表示被转换为体素形式，分割出手部区域后馈送入该网络；之后，通过多次的 3D 卷积、3D 池化和 3D 反卷积等操作，网络从体积表示中提取多尺度 3D 特征，最后回归手部关节点的空间坐标。为了使网络模型对不同的手形大小和镜头视角更具鲁棒性，本文方法对体素形式的手势姿态进行数据增强。整体网络结构如图 1 所示，其中每个模块下方的数字表示输入（上方）和输出（下方）的特征图的“尺寸@通道数量”，其中 N^3 表示 N x N x N。下文将对方法细节进行介绍。

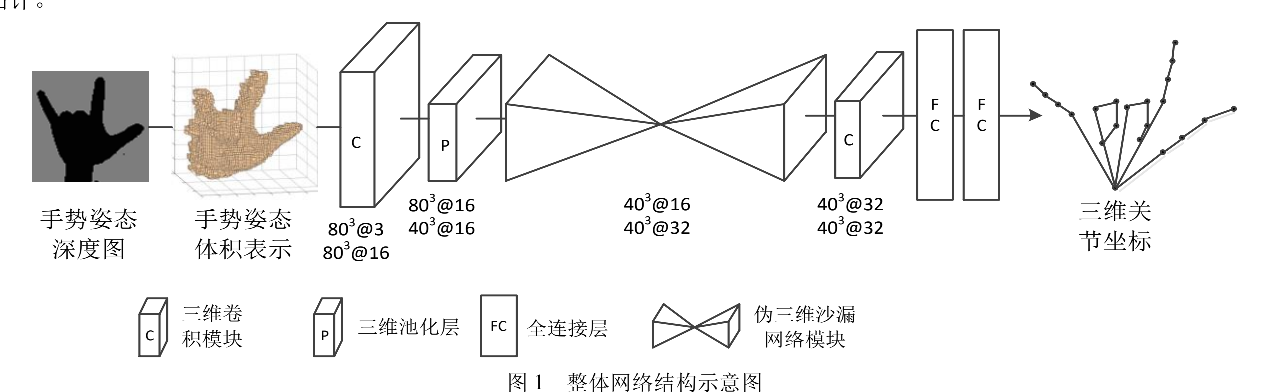


图 1　整体网络结构示意图

Fig. 1　Framework of the proposed method

2.1 手势姿态的体积表示

a)体积表示。将手势姿态编码为体积表示的目标是尽可能的从深度图中表示手势姿态在空间中的 3D 体积。本文工作改进了文献[28]所提出的占用网格模型，采用新的手部区域获取方式。首先，深度图中的每个像素根据深度值重投影到 3D 空间，之后按照预先定义的体素分辨率 v 将该空间分割为体素网格。如式(1)所示，如果某一体素网格中包含有深度点，则将该体素值 $H(i,j,k)$ 设置为 1，否则设为 0。

$$H(\mathrm{i},j,k)=\begin{cases}1 & \left(k-1<\dfrac{P(m,n)}{v}\le k\right)\\ 0 & (\text{otherwise})\end{cases} \tag{1}$$

其中：$H(i,j,k)$ 表示手势姿态体素，$P(m,n)$ 表示深度图中的深

度点，v 表示预先定义的体素分辨率。

为了确定手部区域参考点，本文采用新的基于二维残差网络（ResNet）[29]的浅层网络学习手部中指掌指关节（metacarpophalangeal，MCP）的参考点偏移矢量，相比于文献[24]中采用的浅层 CNN，本文方法可以提取手势深度图更有效的特征，回归手部区域精确参考点，并以该点为中心绘制立体方框，去除无效区域的影响。

b)数据增强。在实际应用场景中，手势姿态在手形大小和观察视角上具有很大的变化，为了使模型更具鲁棒性，本文对训练数据进行数据增强操作。具体来说，分割而来的手部区域将在 XY 空间中随机进行旋转，在 3D 空间中随机执行放缩与平移操作，其中旋转角区间为[-45,45]，放缩因子区间为[0.7,1.3]，平移像素区间为[-10,10]，原始数据集和增强后的数据集均被用做训练数据集。

2.2 网络结构

2.2.1 基础模块

本文提出的伪三维沙漏结构设计可以用来提取不同尺度的特征再进行融合。小尺度上的局部特征对于手部关节点的位置估计至关重要，大尺度上的全局特征则能充分利用手部关节之间的潜在整体相关性，从而提高手势姿态估计的精度。为此，本文主要使用以下三种基础模块来搭建网络模型。

a)三维卷积模块。该模块包含标准三维卷积层、三维批量归一化和激活函数（即 ReLU），主要应用于网络的两端，分别用来对输入进行浅层特征提取，和对整合的各尺度特征进行滤波操作。

b)三维反卷积模块。该模块由三维反卷积层、三维批量归一化和激活函数（即 ReLU）组成，位于沙漏结构的后半段，主要用来对特征图进行上采样操作。

c)伪三维残差模块。ResNet[29]由许多残差块组成，每个残差块可表示为

$$x_{t+1} = x_t + R(x_t) \tag{2}$$

其中: x_t 和 x_{t+1} 表示第 t 个残差单元的输入和输出，R 是非线性残差函数。ResNet 的主要思想是拟合残差函数 R，而不是直接学习映射的非线性函数。

为了减少神经网络模型参数并提取更有效的特征，如图 2 所示，本文采用新的滤波器组合，将原始三维残差模块（图 2(a)）中的三维卷积滤波器（3x3x3 卷积）分解为空间卷积滤波器 S（3x3x1 卷积）和深度卷积滤波器 D（1x1x3 卷积），如图 2(b)，其中 S 用来提取手势姿态的空间特征，D 用来进一步提取手势姿态的深度特征，两种卷积滤波器通过级联的方式组成基础伪三维残差模块，该模块可表示为

$$x_{t+1} = x_t + D(S(x_t)) \tag{3}$$

其中: x_t 和 x_{t+1} 表示第 t 个残差单元的输入和输出，S 和 D 是两个同一路径上的非线性残差函数。

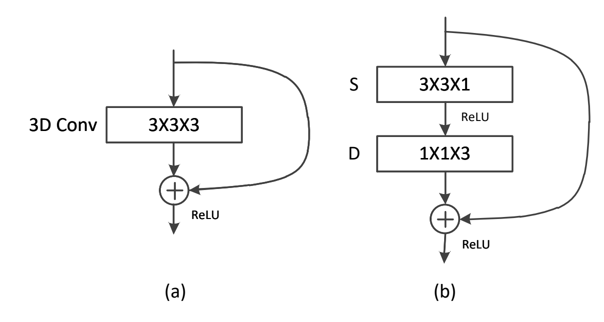


图 2 原始三维残差模块和伪三维残差模块

Fig. 2 Original 3D residual module and pseudo 3D residual module

为了简化模型学习过程，加快模型训练速度，本文在所有基础模块中均添加批量归一化层和激活函数，并将三维卷积模块中的卷积核尺寸设为 3x3x3，3D 池化层和三维反卷积模块中的核尺寸设为 2x2x2，步长设为 2。

2.2.2 伪三维沙漏结构

为了解决现有三维结构网络不能充分利用多尺度特征的问题，本文提出一种新的伪三维沙漏结构，从各个尺度提取特征并进行融合，提高姿态估计精度。体积表示的手势姿态经过多次连续的池化操作获得不同尺度的特征图，伪三维残差模块对多尺度特征图进行特征提取获得特征 $f_k, k \in (1,2,\ldots,n)$；为了整合不同尺度的特征，最小尺寸的特征图多次经过三维反卷积模块进行上采样操作，并融合特征提取阶段获得的 f_k 辅助上采样，最终将各个尺度上的特征整合，充分利用手势姿态的三维信息，确定手部关节点三维坐标。伪三维沙漏结构如图 3 所示，图中每个模块下方的数字表示输入（上方）和输出（下方）的特征图的“尺寸@通道数量”，其中 N3 表示 N x N x N。

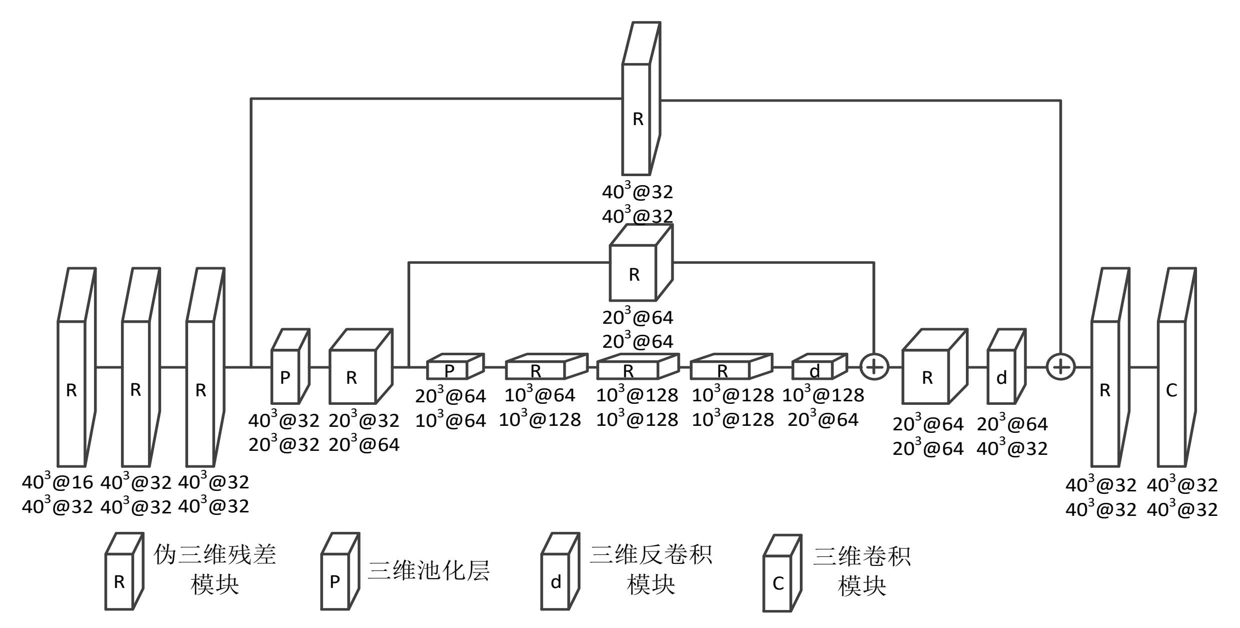


图 3 伪三维沙漏结构

Fig. 3 Pseudo-3D hourglass structure

在沙漏结构中，三维池化层减小特征图的空间尺寸，而伪三维残差模块增加特征图的通道数量，提取手势姿态的三维特征。这些特征将分为两路继续传递，一路会多次进行池化与伪三维卷积操作，分支路则在经过一次简单的滤波操作后，与上采样后的小尺度特征进行融合。在特征图达到最低分辨率时，经过 3 个连续的伪三维残差模块提取三维特征后，使用三维反卷积模块对其进行上采样，再与由分支路而来的特征进行融合。由于沙漏结构是对称的，每次特征融合时均有对应的分支路特征辅助上采样，从而保证沙漏结构整体的输入与输出具有相同的尺寸。

就整体网络结构而言，如图 1 所示，体积表示的手势姿态经过一个卷积核尺寸为 7x7x7 的三维卷积模块滤波处理后，使用三维池化层将尺寸缩小为适合伪三维沙漏结构的输入尺寸，再经过 3 个连续的伪三维残差模块，输入沙漏结构中。达到网络的输出分辨率后，应用 2 个连续的 3x3x3 标准三维卷积处理整合后的特征，再通过 2 个全连接层回归手部关节点的三维坐标。

2.3 网络训练

由于伪三维结构的设计，本文方法大幅降低了网络训练所需的计算成本和空间需求。训练过程中，网络不加载预训练模型，且损失函数 L 采用均方误差计算，如式(4)所示。

$$L=\sum_{m=1}^{M}\sum_{i,j,k}\|C_m^G(i,j,k)-C_m(i,j,k)\|^2 \tag{4}$$

其中：C_m^G 和 C_m 分别是第 m 个手部关节点的真值三维坐标和估计的三维坐标， M 表示单个手部的关节点数量。

该网络使用 Torch7 框架，在单块 NVIDIA Titan X GPU 上进行训练和测试。网络中的所有权重参数均使用 σ=0.001 的零均值高斯分布进行初始化，并由均方根优化算法（root mean square prop，RMSProp）进行更新，学习率设置为 2.5e^{-4}，最小批量为 8。根据 GPU 实际的内存容量，本文将手势姿态的体积大小设为 80 x 80 x 80，为了取得效果最好的模型，每次进行 8 轮训练，共需大约 5 天时间。

3 实验

3.1 手势姿态数据集与评估标准

a)ICVL 手势数据集。ICVL 数据集[30]由含有 33.1 万个深度图的训练集和超过 1500 个深度图的测试集组成，使用英特尔的 Creative Interactive Gesture Camera 从 10 个不同的手势执行人中收集而来。该数据集中每个手指为三个关节点，手掌为一个关节点，共 16 个关节点。

b)NYU 手势数据集。NYU 数据集[31]由含有 7.2 万个深度图的训练集和 8252 个深度图的测试集组成。其中训练集的图像由一人完成，而测试集则由两个人从三个不同视角的 Kinect 生成。该数据集中单手含有 36 个关节点。由于之前的大部分工作仅使用正视图及其中的 14 个关节点进行效果评估，为了便于比较，本文也将按照该配置进行实验。

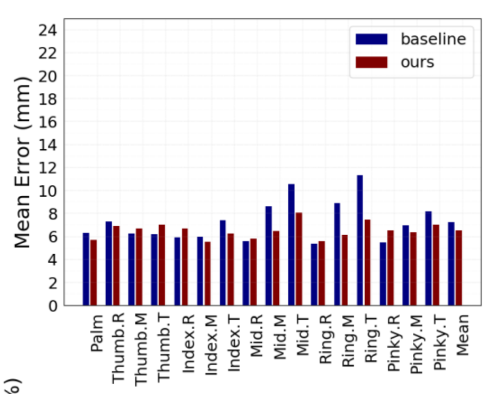
手势估计效果的评估将采用常用的两个评估标准进行，分别为每个关节点的三维距离平均误差（Mean Error）和最大误差低于阈值的正样本图像比例。

3.2 实验结果展示与分析

1)与基准实验的比较

为了探究伪三维沙漏结构对手势估计效果是否有提升，网络模型在 ICVL 数据集[30]上进行了比较实验。在该实验中，仅将沙漏结构网络从二维扩展至三维，即三维沙漏网络，将其作为基准实验与本文所提出的伪三维沙漏结构网络进行比较，手势估计效果比较结果如图 4 所示。此外，为了探究伪三维沙漏结构对模型尺寸的减小和手势姿态估计速度是否有所提高，还比较了本文方法与基准实验之间的模型大小和单次手势姿态估计的时间，比较结果如表 1 所示。

从实验结果可以看出，在误差阈值较小时，本文方法中的准确率略高于基准实验，同时三维距离平均误差比基准实验要低 0.754mm，单次手势估计时间加快了 4ms，模型尺寸大幅减小为基准实验的 1/2。可见伪三维沙漏网络在小幅提升手势估计精度的同时，能显著加快手势估计速度，并大幅减小模型尺寸。



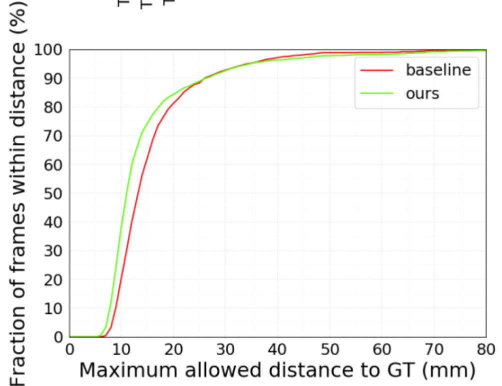


图 4 ICVL 数据集上的基准实验比较结果

Fig. 4 Comparison of the proposed method with baseline on ICVL dataset

表 1 基准实验模型与本文方法模型比较实验结果

Table 1 Comparison of the proposed method with baseline

方法	Mean Error(mm)	模型大小	单次手势姿态估计时间（ms）
基准实验	7.275	26.1MB	12.27
本文方法	6.521	12.5MB	8.39

2)与其他方法的比较

该实验在 ICVL[30]和 NYU[31]两个数据集上进行本文方法与多种方法之间的比较，其中包括潜在随机森林（latent random forest，LRF）[20]，级联手势回归（cascade）[21]，改进 DeepPrior（DeepPrior++）[24]，Hand3D[14]，CrossingNets[25]，姿态导向区域集合网络（Pose-REN）[24]，密集 3D 回归方法（DenseReg）[13]，反馈环训练方法（Feedback）[32]和基于 3D CNN 的方法（3D CNN）[1]。以上方法的结果均根据线上所提供的预测标签计算而来。

如图 5、6 和表 2 所示，在 ICVL 数据集[30]上本文方法要优于先前的所有方法，在误差允许范围较小的情况下，仍有良好表现。而在 NYU 数据集[31]上，本文方法的精度要略微低于 DenseReg[13]。造成这种现象的原因可能是由于在该数据集中手部区域没有被裁剪出来，所以在对区域进行分割时产生误差。尽管如此，按照文献[13]中介绍 DenseReg 方法进行单张手势姿态估计需要 36 ms，而本文方法仅需要 8.39ms，运行速度远快于 DenseReg 方法。

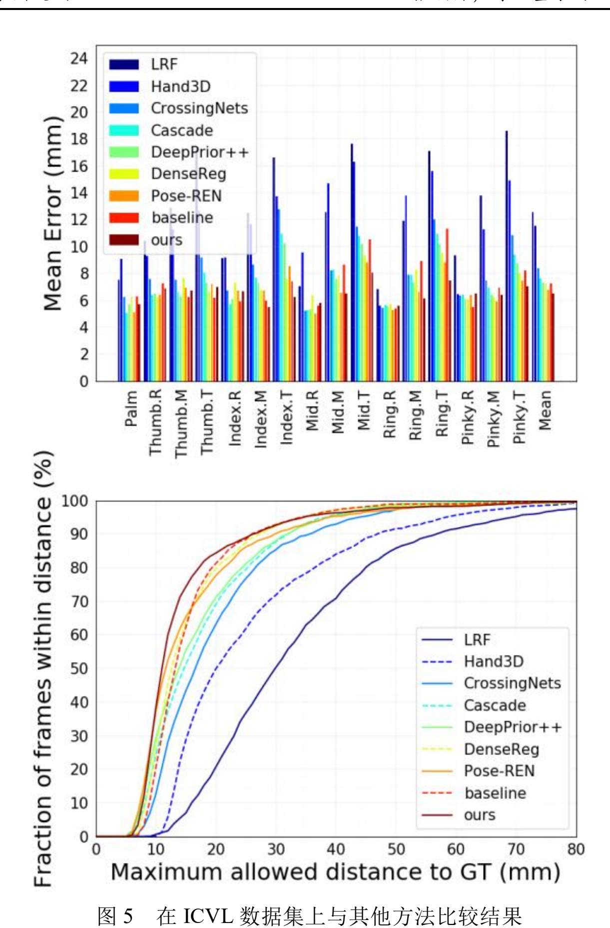
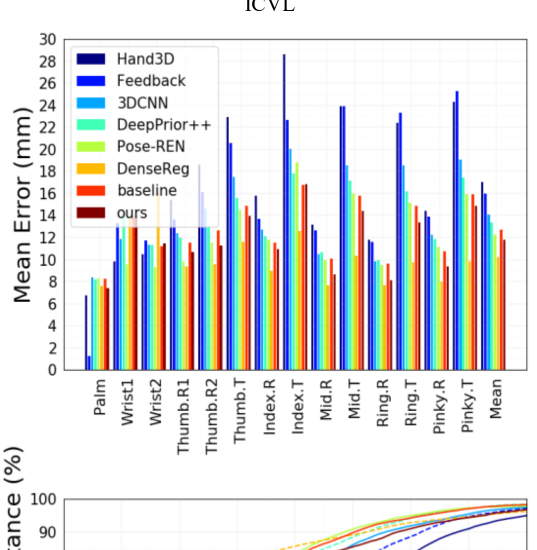


图 5　在 ICVL 数据集上与其他方法比较结果

Fig. 5　Comparison of the proposed method with other methods on ICVL



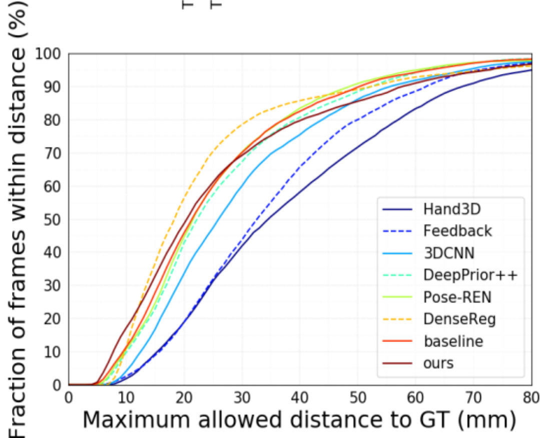


图 6　在 NYU 数据集上与其他方法比较结果

Fig. 6　Comparison of the proposed method with other methods on NYU

表 2　在 ICVL 和 NYU 上各种方法的三维距离平均误差比较结果

Table 2　Comparison of 3D distance mean error of various methods on ICVL and NYU

(a) ICVL

方法	**Mean Error(mm)**
LRF	12.58
Hand3D	10.9
CrossingNets	10.2
Cascade	9.9
DeepPrior++	8.1
DenseReg	7.24
Pose-REN	6.79
基准实验	7.28
本文方法	**6.52**

(b) NYU

方法	**Mean Error(mm)**
Hand3D	17.6
Feedback	15.97
3DCNN	14.1
DeepPrior++	12.24
Pose-REN	11.81
DenseReg	**10.21**
基准实验	12.14
本文方法	11.31

4　结束语

本文提出了一种精确的基于伪三维卷积神经网络的手势姿态估计方法。手势深度图编码为三维体积表示后，作为伪三维卷积神经网络的输入，并使用改进的分割方法对手部区域进行分割。通过使用空间卷积滤波器和深度卷积滤波器级联的方式，简化了标准三维卷积。在多尺度下进行特征的提取与融合，使手势姿态中的三维信息得到充分的利用。实验结果表明，本文方法中的模型具有较小的尺寸，在提高精度的同时，加快了手势姿态估计的速度。在未来的工作中，将会进一步探究多样性的网络结构对手势姿态估计效果的影响。

参考文献：

[1] Ge Liuhao, Liang Hui, Yuan Junsong, *et al.* 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2017: 5679-5688.

[2] Yuan Shanxin, Ye Qi, Stenger Bjorn, *et al.* BigHand2. 2M benchmark: hand pose dataset and state of the art analysis [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2017: 35-45,

[3] Shotton J, Kipman A, Blake A, *et al.* Efficient human pose estimation from single depth images [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2013, 35 (12): 2821-2840.

[4] Qian Chen, Sun Xiaoli, Wei Yichen, *et al.* Realtime and Robust Hand Tracking from Depth [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2014: 1106-1113.

[5] 徐岳峰，周书仁，王刚，等. 基于深度图像梯度特征的人体姿态估计 [J]. 计算机工程, 2015, 41 (12): 200-205. (Xu Yuefeng, Zhou Shuren, Wang Gang, *et al.* Human body attitude estimation based on gradient

feature of depth images [J]. Computer Engineering, 2015, 41 (12): 200-205.)

[6] 王松，刘复昌，黄骥，等. 基于卷积神经网络的深度图姿态估计算法研究 [J]. 系统仿真学报, 2017, 29 (11): 2618-2623. (Wang Song, Liu Fuchang, Huang Ji, *et al.* Pose estimation using convolutional neural network with synthesis depth data [J]. Joural of System Simulation, 2017, 29 (11): 2618-2623.)

[7] Tompson J, Stein M, Lecun Y, *et al.* Real-time continuous pose recovery of human hands using convolutional networks [J]. ACM Trans on Graphics, 2014, 33 (5): 1-10.

[8] Oberweger M, Wohlhart P, Lepetit V. Hands deep in deep learning for hand pose estimation [EB/OL].(2015-10-02). https://www.tugraz.at/.../3d_hand_pose/cvww15_presentation.pdf.

[9] Ge Liuhao, Liang Hui, Yuan Junsong, *et al.* Robust 3D Hand Pose Estimation in Single Depth Images: from Single-View CNN to Multi-View CNNs [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2016: 3593-3601.

[10] Sinha A, Choi C, Ramani K. DeepHand: robust hand pose estimation by completing a matrix imputed with deep features [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2016: 4150-4158.

[11] Zhou Xingyi, Wan Qingfu, Zhang Wei, *et al.* Model-based Deep Hand Pose Estimation [C]//Proc of the 25th International Joint Conference on Artificial Intelligence.2016: 2421-2427.

[12] Tompson J, Stein M, Lecun Y, *et al.* Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks [J]. ACM Trans on Graphics, 2014, 33 (5): 1-10.

[13] Wan Chengde, Probst T, Van Gool L, *et al.* Dense 3D regression for hand pose estimation [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2018.

[14] Deng Xiaoming, Yang Shuo, Zhang Yinda, *et al.* Hand3D: hand pose estimation using 3D neural network [EB/OL].(2017). http://cn. arxiv. org/pdf/1704. 02224.

[15] Qiu Zhaofan, Yao Ting, Mei Tao. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2017: 5534-5542.

[16] Newell A, Yang Kaiyu, Deng Jia. Stacked hourglass networks for human pose estimation [C]//Proc of European Conference on Computer Vision. 2016: 483-499.

[17] Tagliasacchi A, Tkach A, Bouaziz S, *et al.* Robust articulated-ICP for real-time hand tracking [C]//Proc of Eurographics Symposium on Geometry Processing. 2015: 101-114.

[18] Oikonomidis I, Kyriazis N, Argyros A. Efficient model-based 3D tracking of hand articulations using Kinect [C]//Proc of British Machine Vision Conference. 2011.

[19] Shotton J, Fitzgibbon A, Cook M, *et al.* Real-time human pose recognition in parts from single depth images [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2011: 1297-1304.

[20] Tang Danhang, Chang Hyung Jin, Tejani A, *et al.* Latent regression forest: structured estimation of 3D articulated hand posture [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society, 2014: 3786-3793.

[21] Sun Xiaoli, Wei Yichen, Liang Shuang, *et al.* Cascaded hand pose regression [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2015: 824-832.

[22] Guo Hengkai, Wang Guijin, Chen Xinghao, *et al.* Region ensemble network: improving convolutional network for hand pose estimation [C]// Proc of IEEE International Conference on Image Processing. 2017.

[23] Chen Xinghao, Wang Guijin, Guo Hengkai, *et al.* Pose guided structured region ensemble network for cascaded hand pose estimation [OL]. (2017-0 3).http://cn. arxiv. org/pdf/1708. 03416..

[24] Oberweger M, Lepetit V. DeepPrior+: improving fast and accurate 3D hand pose estimation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society,2017.

[25] Wan Chengde, Probst T, Van Gool L, *et al.* Crossing nets: combining GANs and VAEs with a shared latent space for hand pose estimation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society,2017: 1196-1205.

[26] Wu Zhirong, Song Shuran, Khosla A, *et al.* 3D ShapeNets: A deep representation for volumetric shapes [C]// IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2015: 1912-1920.

[27] Song Shuran, Xiao Jianxiong. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society, 2016: 808-816.

[28] Maturana D, Scherer S. VoxNet: a 3D convolutional neural network for real-time object recognition [C]//Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2015: 922-928.

[29] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep Residual Learning for Image Recognition [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition.Washington DC:IEEE Computer Society, 2016: 770-778.

[30] Tang Danhang, Chang Hyung Jin, Tejani A, *et al.* Latent regression forest: structured estimation of 3D articulated hand posture [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC:IEEE Computer Society, 2014: 3786-3793.

[31] Tompson J, Stein M, Lecun Y, *et al.* Real-Time Continuous pose recovery of human hands using convolutional networks [J]. ACM Trans on Graphics, 2014, 33 (5): 1-10.

[32] Oberweger M, Wohlhart P, Lepetit V. Training a feedback loop for hand pose estimation [C]//Proc of International Conference on Computer Vision. Washington DC:IEEE Computer Society, 2016.